Topic 6: - Topic Analysis

Clarissa Boyajian

2022-05-09

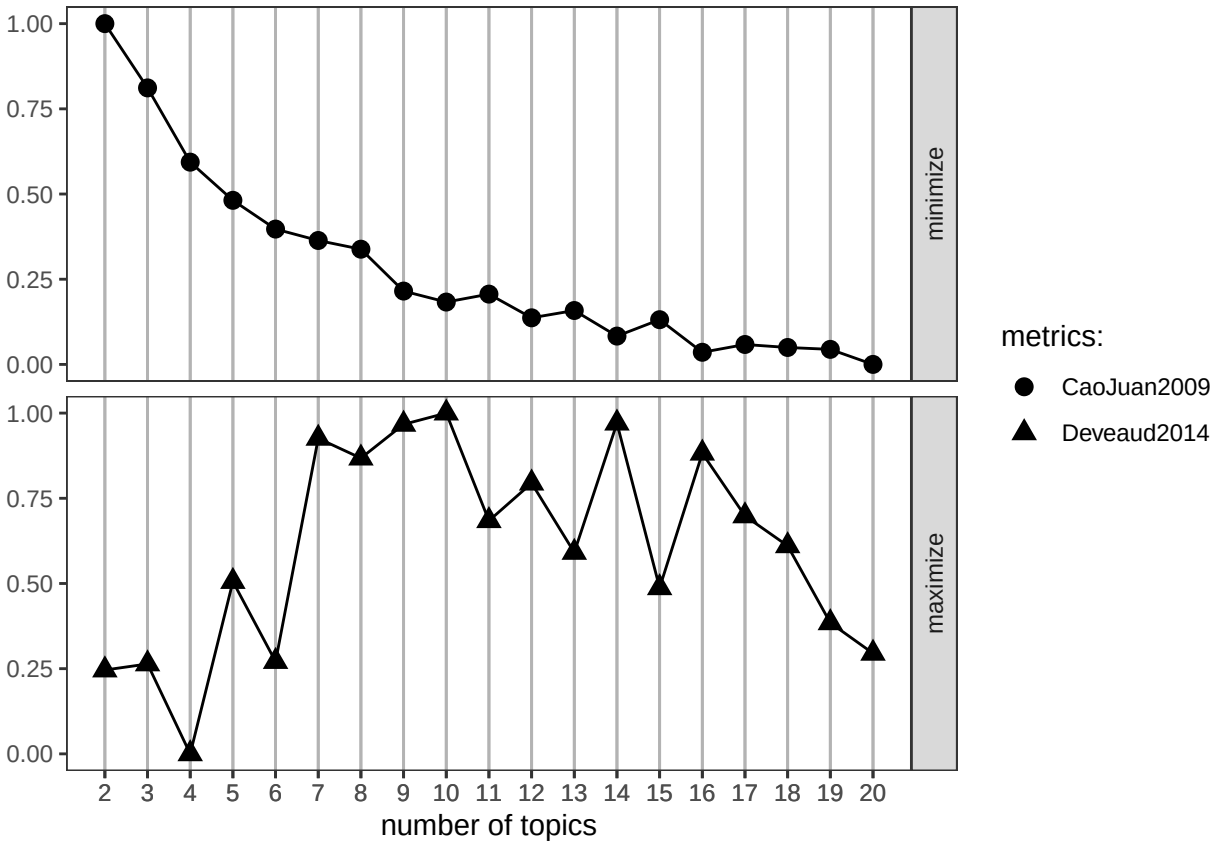
```
## -- read in, clean, and wrangle data -- ##
comments_df<-read_csv(here("data/comments_df.csv"))

epa_corp <- corpus(x = comments_df, text_field = "text")
epa_corp.stats <- summary(epa_corp)
tokens <- tokens(epa_corp, remove_punct = TRUE, remove_numbers = TRUE)

# I added some project-specific stop words here
add_stops <- c(stopwords("en"),"environmental", "justice", "ej", "epa", "public", "comment")
tokens_1 <- tokens_select(tokens, pattern = add_stops, selection = "remove")

doc_feature_matric_common <- dfm(tokens_1, tolower = TRUE)
doc_feature_matric <- dfm_wordstem(doc_feature_matric_common)
doc_feature_matric <- dfm_trim(doc_feature_matric, min_docfreq = 2) #remove terms only appearing in one

# remove rows (docs) with all zeros
select_idx <- slam::row_sums(doc_feature_matric) > 0
doc_feature_matric <- doc_feature_matric[select_idx, ]
```

Calculate value of k that is most likely

```
# calculate what initial value of k is most likely
result <- FindTopicsNumber(
  doc_feature_matric,
  topics = seq(from = 2, to = 20, by = 1),
  metrics = c("CaoJuan2009",  "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 77),
  verbose = TRUE
)
```

```
## fit models... done.
## calculate metrics:
##   CaoJuan2009... done.
##   Deveaud2014... done.
```

```
FindTopicsNumber_plot(result)
```

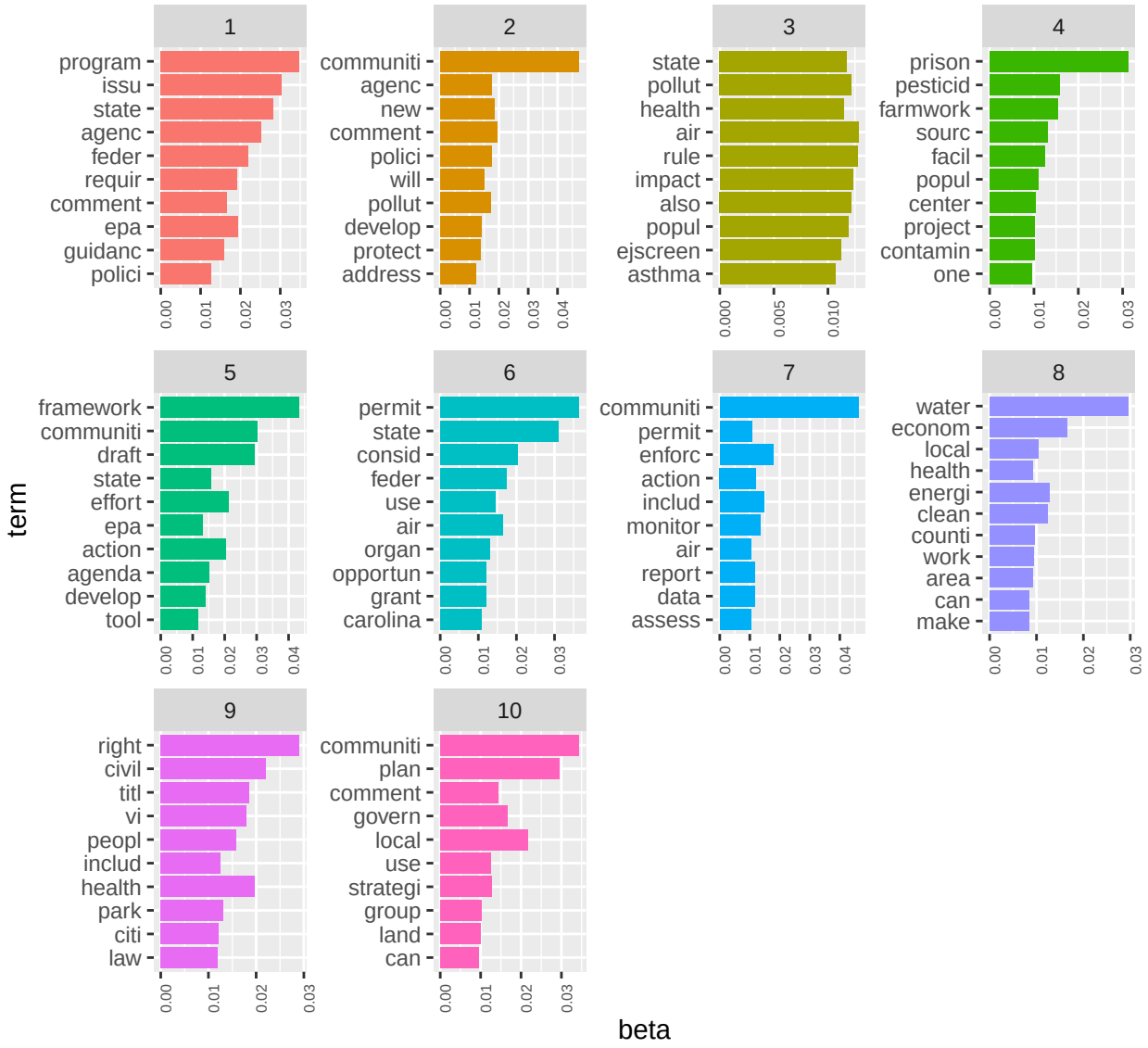


Model 1 (k = 10)

```
k <- 10

topicModel_k10 <- LDA(doc_feature_matric, k,
                      method = "Gibbs",
                      control = list(iter = 500, verbose = 25))

tmResult_k10 <- posterior(topicModel_k10)
theta_k10 <- tmResult_k10$topics
beta_k10 <- tmResult_k10$terms # probability of each term in each topic
vocab_k10 <- (colnames(beta_k10))

comment_topic_k10 <- tidy(topicModel_k10, matrix = "beta")

top_terms_k10 <- comment_topic_k10 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
```

```
top_terms_k10 %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
```

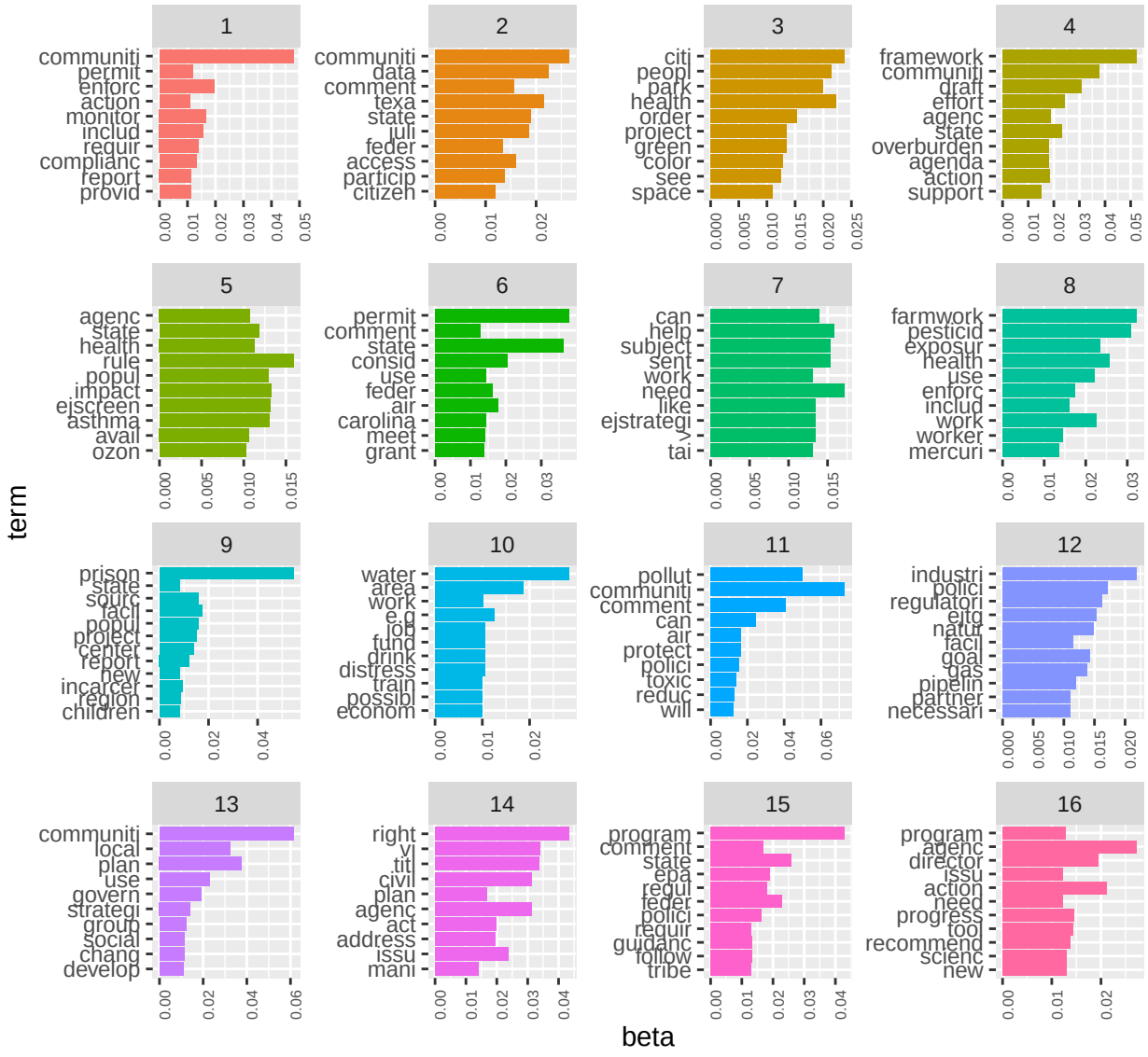
```
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 6))
```



Model 2 (k = 16)

```
k <- 16

topicModel_k16 <- LDA(doc_feature_matric, k,
                      method = "Gibbs",
                      control = list(iter = 500, verbose = 25))

tmResult_k16 <- posterior(topicModel_k16)
theta_k16 <- tmResult_k16$topics
beta_k16 <- tmResult_k16$terms # probability of each term in each topic
vocab_k16 <- (colnames(beta_k16))

comment_topic_k16 <- tidy(topicModel_k16, matrix = "beta")

top_terms_k16 <- comment_topic_k16 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
```

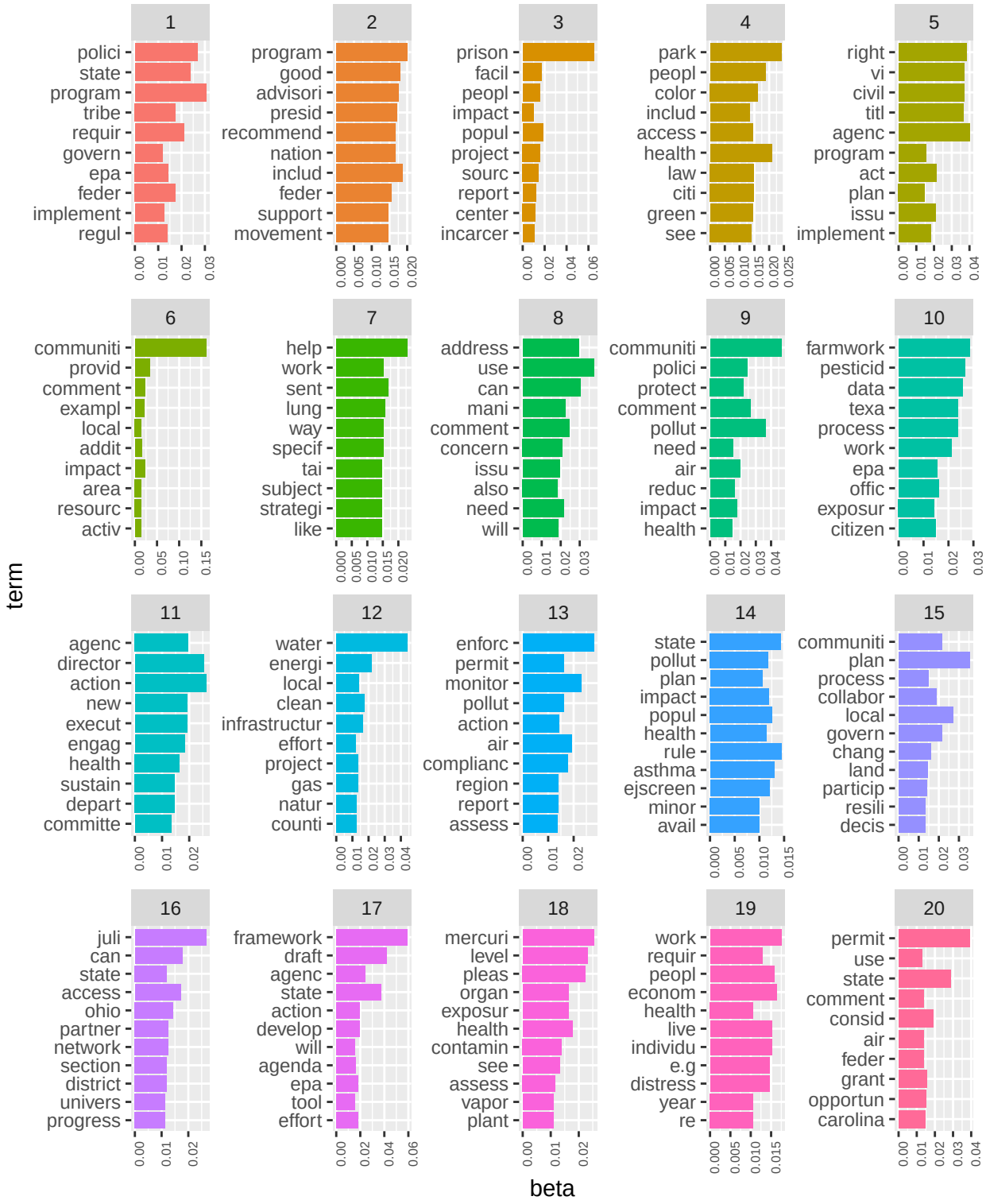
```
top_terms_k16 %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 6))
```


1
communiti
permit
enforc
action
monitor
includ
requir
complianc
report
provid
0.00
0.01
0.02
0.03
0.04
0.05
2
communiti
data
comment
texa
state
juli
feder
access
particip
citizen
0.00
0.01
0.02
3
citi
peopl
park
health
order
project
green
color
see
space
0.000
0.005
0.010
0.015
0.020
0.025
4
framework
communiti
draft
effort
agenc
state
overburden
agenda
action
support
0.00
0.01
0.02
0.03
0.04
0.05
5
agenc
state
health
rule
popul
impact
ejscreen
asthma
avail
ozon
0.000
0.005
0.010
0.015
6
permit
comment
state
consid
use
feder
air
carolina
meet
grant
0.00
0.01
0.02
0.03
7
can
help
subject
sent
work
need
like
ejstrategi
tai
0.000
0.005
0.010
0.015
8
farmwork
pesticid
exposur
health
use
enforc
includ
work
worker
mercuri
0.00
0.01
0.02
0.03
9
prison
state
sourc
facil
popul
project
center
report
new
incarcer
region
children
0.00
0.02
0.04
10
water
area
work
e.g
job
fund
drink
distress
train
possibl
econom
0.00
0.01
0.02
11
pollut
communiti
comment
can
air
protect
polici
toxic
reduc
will
0.00
0.02
0.04
0.06
12
industri
polici
regulatori
ejtg
natur
facil
goal
gas
pipelin
partner
necessari
0.000
0.005
0.010
0.015
0.020
13
communiti
local
plan
use
govern
strategi
group
social
chang
develop
0.00
0.02
0.04
0.06
14
right
vi
tit
civil
plan
agenc
act
address
issu
mani
0.00
0.01
0.02
0.03
0.04
15
program
comment
state
epa
regul
feder
polici
requir
guidanc
follow
tribe
0.00
0.01
0.02
0.03
0.04
16
program
agenc
director
issu
action
need
progress
tool
recommend
scienc
new
0.00
0.01
0.02
term
beta

Model 3 (k = 20)

```
k <- 20

topicModel_k20 <- LDA(doc_feature_matric, k,
                      method = "Gibbs",
                      control = list(iter = 500, verbose = 25))

tmResult_k20 <- posterior(topicModel_k20)
theta_k20 <- tmResult_k20$topics
beta_k20 <- tmResult_k20$terms # probability of each term in each topic
vocab_k20 <- (colnames(beta_k20))

comment_topic_k20 <- tidy(topicModel_k20, matrix = "beta")

top_terms_k20 <- comment_topic_k20 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
```

```
top_terms_k20 %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 6))
```


1
polici
state
program
tribe
requir
govern
epa
feder
implement
regul
2
program
good
advisori
presid
recommend
nation
includ
feder
support
movement
3
prison
facil
peopl
impact
popul
project
sourc
report
center
incarcer
4
park
peopl
color
includ
access
health
law
citi
green
see
5
right
vi
civil
titl
agenc
program
act
plan
issu
implement
6
communiti
provid
comment
exampl
local
addit
impact
area
resourc
activ
7
help
work
sent
lung
way
specif
tai
subject
strategi
like
8
address
use
can
mani
comment
concern
issu
also
need
will
9
communiti
polici
protect
comment
pollut
need
air
reduc
impact
health
10
farmwork
pesticid
data
texa
process
work
epa
offic
exposur
citizen
11
agenc
director
action
new
execut
engag
health
sustain
depart
committe
12
water
energi
local
clean
infrastructur
effort
project
gas
natur
counti
13
enforc
permit
monitor
pollut
action
air
complianc
region
report
assess
14
state
pollut
plan
impact
popul
health
rule
asthma
ejscreen
minor
avail
15
communiti
plan
process
collabor
local
govern
chang
land
particip
resili
decis
16
juli
can
state
access
ohio
partner
network
section
district
univers
progress
17
framework
draft
agenc
state
action
develop
will
agenda
epa
tool
effort
18
mercuri
level
pleas
organ
exposur
health
contamin
see
assess
vapor
plant
19
work
requir
peopl
econom
health
live
individu
e.g
distress
year
re
20
permit
use
state
comment
consid
air
feder
grant
opportun
carolina
term
beta

Pick best k value

```
docIds <- c(1, 2, 3, 4, 5, 6)
N <- length(docIds)
```

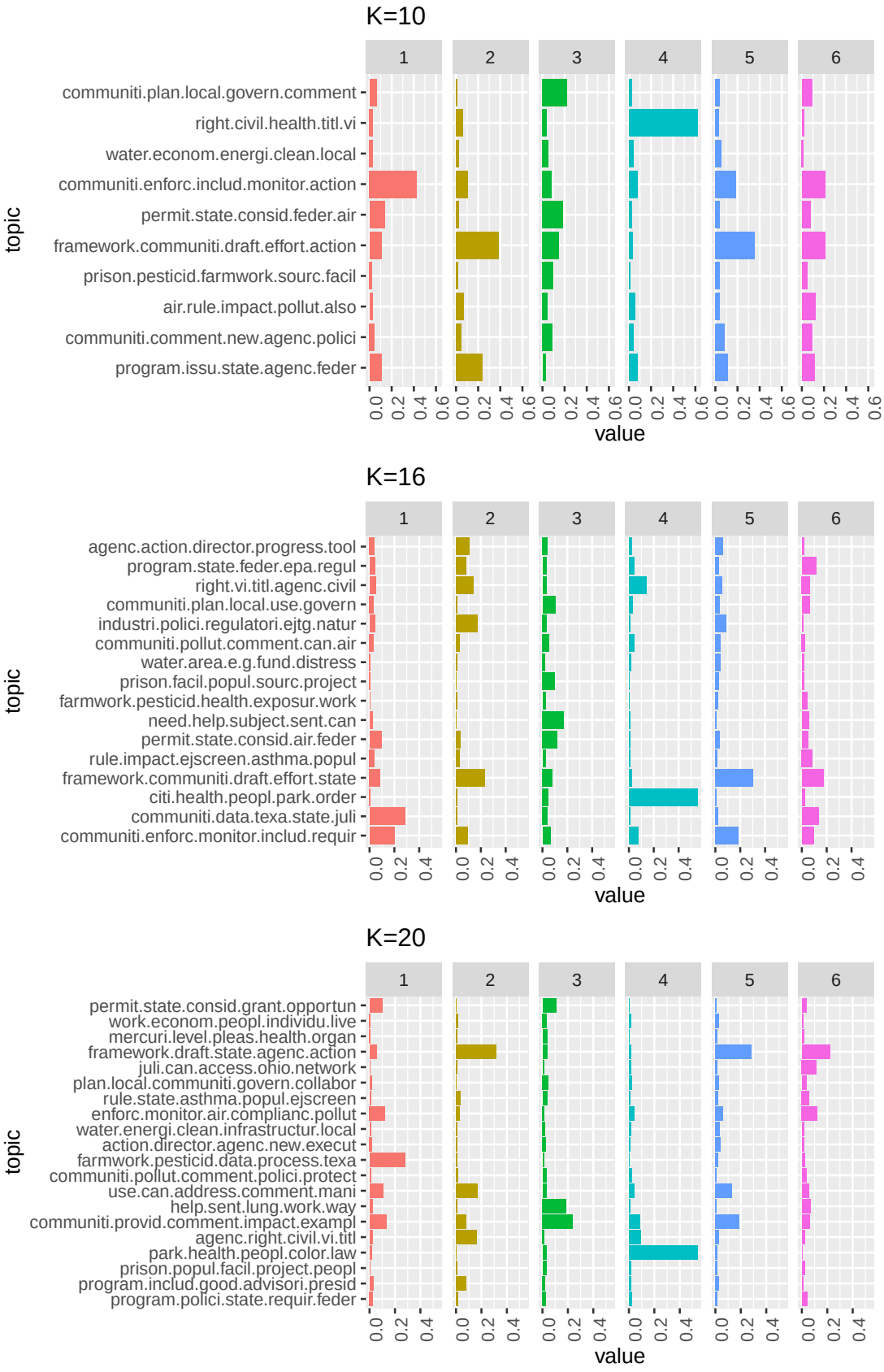
```
## -- k = 10 -- ##
# name each topic based on 1st 5 terms in topic
top5termsPerTopic_k10 <- terms(topicModel_k10, 5)
topicNames_k10 <- apply(top5termsPerTopic_k10, 2, paste, collapse = " ")

# get topic proportions from example documents
topicProportions_K10 <- theta_k10[docIds,]
colnames(topicProportions_K10) <- topicNames_k10

vizDataFrame_k10 <- melt(cbind(data.frame(topicProportions_K10),
                           document = factor(1:N)),
                     variable.name = "topic",
                     id.vars = "document")

plot_k10 <- ggplot(data = vizDataFrame_k10,
                   aes(x = topic,
                       y = value,
                       fill = document),
                   ylab = "proportion") +
  geom_bar(stat="identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = "none") +
  coord_flip() +
  facet_wrap(~ document, ncol = N) +
  labs(title = "K=10")
```

```
## -- k = 16 -- ##
# name each topic based on 1st 5 terms in topic
top5termsPerTopic_k16 <- terms(topicModel_k16, 5)
topicNames_k16 <- apply(top5termsPerTopic_k16, 2, paste, collapse = " ")

# get topic proportions from example documents
topicProportions_k16 <- theta_k16[docIds,]
colnames(topicProportions_k16) <- topicNames_k16

vizDataFrame_k16 <- melt(cbind(data.frame(topicProportions_k16),
                           document = factor(1:N)),
                     variable.name = "topic",
                     id.vars = "document")

plot_k16 <- ggplot(data = vizDataFrame_k16,
                   aes(x = topic,
                       y = value,
                       fill = document),
                   ylab = "proportion") +
  geom_bar(stat="identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = "none") +
```

```
  coord_flip() +
  facet_wrap(~ document, ncol = N) +
  labs(title = "K=16")
```

```
## -- k = 20 -- ##
# name each topic based on 1st 5 terms in topic
top5termsPerTopic_k20 <- terms(topicModel_k20, 5)
topicNames_k20 <- apply(top5termsPerTopic_k20, 2, paste, collapse = " ")

# get topic proportions from example documents
topicProportions_k20 <- theta_k20[docIds,]
colnames(topicProportions_k20) <- topicNames_k20

vizDataFrame_k20 <- melt(cbind(data.frame(topicProportions_k20),
                               document = factor(1:N)),
                         variable.name = "topic",
                         id.vars = "document")

plot_k20 <- ggplot(data = vizDataFrame_k20,
                   aes(x = topic,
                       y = value,
                       fill = document),
                   ylab = "proportion") +
  geom_bar(stat="identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = "none") +
  coord_flip() +
  facet_wrap(~ document, ncol = N) +
  labs(title = "K=20")
```

```
plot_k10 / plot_k16 / plot_k20
```


K=10
1
2
3
4
5
6
communiti.plan.local.govern.comment
right.civil.health.titl.vi
water.econom.energi.clean.local
communiti.enforc.includ.monitor.action
permit.state.consid.feder.air
framework.communiti.draft.effort.action
prison.pesticid.farmwork.sourc.facil
air.rule.impact.pollut.also
communiti.comment.new.agenc.polici
program.issu.state.agenc.feder
topic
0.0
0.2
0.4
0.6
value
K=16
agenc.action.director.progress.tool
program.state.feder.epa.regul
right.vi.titl.agenc.civil
communiti.plan.local.use.govern
industri.polici.regulatori.ejtg.natur
communiti.pollut.comment.can.air
water.area.e.g.fund.distress
prison.facil.popul.sourc.project
farmwork.pesticid.health.exposur.work
need.help.subject.sent.can
permit.state.consid.air.feder
rule.impact.ejscreen.asthma.popul
framework.communiti.draft.effort.state
citi.health.peopl.park.order
communiti.data.texa.state.juli
communiti.enforc.monitor.includ.requir
topic
0.0
0.2
0.4
value
K=20
permit.state.consid.grant.opportun
work.econom.peopl.individu.live
mercuri.level.pleas.health.organ
framework.draft.state.agenc.action
juli.can.access.ohio.network
plan.local.communiti.govern.collabor
rule.state.asthma.popul.ejscreen
enforc.monitor.air.complianc.pollut
water.energi.clean.infrastructur.local
action.director.agenc.new.execut
farmwork.pesticid.data.process.texa
communiti.pollut.comment.polici.protect
use.can.address.comment.mani
help.sent.lung.work.way
communiti.provid.comment.impact.exampl
agenc.right.civil.vi.titl
park.health.peopl.color.law
prison.popul.facil.project.peopl
program.includ.good.advisori.presid
program.polici.state.requir.feder
topic
0.0
0.2
0.4
value

```
svd_tsne <- function(x){
  tsne(svd(x)$u)
  }
```

```
json_k10 <- createJSON(
  phi = tmResult_k10$terms,
  theta = tmResult_k10$topics,
  doc.length = rowSums(doc_feature_matric),
  vocab = colnames(doc_feature_matric),
  term.frequency = colSums(doc_feature_matric),
  mds.method = svd_tsne,
  plot.opts = list(xlab = "", ylab = ""))

json_k16 <- createJSON(
  phi = tmResult_k16$terms,
  theta = tmResult_k16$topics,
  doc.length = rowSums(doc_feature_matric),
  vocab = colnames(doc_feature_matric),
  term.frequency = colSums(doc_feature_matric),
  mds.method = svd_tsne,
  plot.opts = list(xlab = "", ylab = ""))

json_k20 <- createJSON(
  phi = tmResult_k20$terms,
  theta = tmResult_k20$topics,
  doc.length = rowSums(doc_feature_matric),
  vocab = colnames(doc_feature_matric),
  term.frequency = colSums(doc_feature_matric),
  mds.method = svd_tsne,
  plot.opts = list(xlab = "", ylab = ""))

serVis(json_k20)
```

Based on the above minimization and maximization plots, 10 was the plot with the highest maximization value. We know there are 9 priority areas for the EPA. So it would make sense that there are topics for each priority area, plus one extra topic that catches on additional miscellaneous topic that appears in all PDFs.